

SPECIFICATION

Electronic Version 1.2.8

Stylesheet Version 1.0

METHOD, SYSTEM AND COMPUTER SOFTWARE FOR PROVIDING A GENOMIC WEB PORTAL

Cross Reference to Related Applications

The present application claims priority from U.S. Provisional Patent Application Serial No. 60/288,429, entitled "METHOD, SYSTEM, AND COMPUTER SOFTWARE FOR PROVIDING A GENOMIC WEB PORTAL," filed May 3, 2001; U.S. Provisional Patent Application, Serial No. 60/301,298, entitled "WEB APPLICATION FOR DESIGNING AND ORDERING FLEXIBLE CONTENT", filed June 25, 2001; U.S. Provisional Patent Application, Serial No. 60/306,033, entitled "PROBESET ANNOTATIONS", filed July 16, 2001; U.S. Provisional Patent Application Serial No. 60/333,522, entitled "METHOD, SYSTEM, AND COMPUTER SOFTWARE FOR PROVIDING A GENOMIC WEB PORTAL," filed November 27, 2001; U.S. Provisional Patent Application, Serial No. 60/343,511, entitled "METHOD, SYSTEM, AND COMPUTER SOFTWARE FOR PROVIDING A GENOMIC WEB PORTAL," filed December 21, 2001; U.S. Provisional Patent Application Serial No. 60/349,546, entitled "METHOD, SYSTEM, AND COMPUTER SOFTWARE FOR PROVIDING A GENOMIC WEB PORTAL", filed January 18, 2002; and U.S. Provisional Patent Application Attorney Docket No. 3291.6, entitled "METHOD, SYSTEM, AND COMPUTER SOFTWARE FOR PROVIDING A GENOMIC WEB PORTAL", filed April 26, 2002, all of which are hereby incorporated herein by reference in their entireties for all purposes. The present application is also a continuation in part of, and claims priority from, U.S. Provisional Patent Application Serial No. 60/178,077, entitled "METHOD, SYSTEM, AND COMPUTER SOFTWARE FOR PROVIDING A GENOMIC WEB PORTAL," filed January 25, 2000; Patent Cooperation Treaty Application Number PCT/US 01/02316 entitled

"METHOD, SYSTEM, AND COMPUTER SOFTWARE FOR PROVIDING A GENOMIC WEB PORTAL," filed January 24, 2001.

Background of Invention

- [0001] Field of the Invention: The present invention relates to the field of bioinformatics. In particular, the present invention relates to computer systems, methods, and products for providing genomic information over networks such as the Internet.
- [0002] Related Art: Research in molecular biology, biochemistry, and many related health fields increasingly requires organization and analysis of complex data generated by new experimental techniques. These tasks are addressed by the rapidly evolving field of bioinformatics. See, e.g., H. Rashidi and K. Buehler, *Bioinformatics Basics: Applications in Biological Science and Medicine* (CRC Press, London, 2000); *Bioinformatics: Practical Guide to the Analysis of Gene and Proteins* (B.F. Ouellette and A.D. Bzevanis, eds., Wiley & Sons, Inc.; 2d ed., 2001), both of which are hereby incorporated herein by reference in their entireties. Broadly, one area of bioinformatics applies computational techniques to large genomic databases, often distributed over and accessed through networks such as the Internet, for the purpose of illuminating relationships among gene structure and/or location, protein function, and metabolic processes.

Summary of Invention

- [0003] The expanding use of microarray technology is one of the forces driving the development of bioinformatics. In particular, microarrays and associated instrumentation and computer systems have been developed for rapid and large-scale collection of data about the expression of genes or expressed sequence tags (EST's) in tissue samples. The data may be used, among other things, to study genetic characteristics and to detect mutations relevant to genetic and other diseases or conditions. More specifically, the data gained through microarray experiments is valuable to researchers because, among other reasons, many disease states can potentially be characterized by differences in the expression levels of various genes, either through changes in the copy number of the genetic DNA or through changes in levels of transcription (e.g., through control of initiation, provision of RNA precursors, or RNA processing) of particular genes. Thus, for example, researchers use

microarrays to answer questions such as: Which genes are expressed in cells of a malignant tumor but not expressed in either healthy tissue or tissue treated according to a particular regime? Which genes or EST's are expressed in particular organs but not in others? Which genes or EST's are expressed in particular species but not in others? How does the environment, drugs, or other factors influence gene expression? Data collection is only an initial step, however, in answering these and other questions. Researchers are increasingly challenged to extract biologically meaningful information from the vast amounts of data generated by microarray technologies, and to design follow-on experiments. A need exists to provide researchers with improved tools and information to perform these tasks.

- [0004] Systems, methods, and computer program products are described herein to address these and other needs. In some embodiments, a web portal processes inquiries regarding biological information, biological devices or substances, reagents, and other information or products related to results of microarray experiments. In some implementations, the user selects "probe-set identifiers"(a broad term that is described below) that may be associated with probe sets of one or more probes. These probe sets are capable of enabling detection of biological molecules. These biological molecules include, but are not limited to, nucleic acids including DNA representations or mRNA transcripts and/or representations of corresponding genes (such nucleic acids may hereafter, for convenience, be referred to simply as "mRNA transcripts"). The corresponding genes or EST's are identified and are correlated with related data and/or products, which are provided to the user.
- [0005] In accordance with a particular embodiment, a system is described for providing information related to one or more probe sets. The system includes an input manager that receives from a user a selection of one or more probe-set identifiers associated with one or more probe sets. The probe-set identifiers include one or more annotation terms. The system also includes a determiner that correlates the user-selected probe-set identifiers with one or more biological sequences, and a correlator that correlates the one or more biological sequences with a first set of probe sets. In some implementations the system could also include an output manager that identifies the first set of probe sets to the user. The first set of probe sets may be identified in a tabular format. The one or more probe sets may include probes of a

synthesized probe array and/or a spotted probe array. The biological sequences may include one or more of a whole or a part of a consensus, EST, gene, or SIF sequence. In some implementations, the annotation terms are descriptive of any one or more of molecular function, cellular location, tissue type, or biological process.

- [0006] In accordance with another embodiment, a method is described for providing information related to one or more probe sets. The method includes the acts of receiving from a user a selection of one or more probe-set identifiers associated with one or more probe sets, wherein the probe-set identifiers include one or more annotation terms; correlating the user-selected probe-set identifiers with one or more biological sequences; and correlating the one or more biological sequences with a first set of probe sets.
- [0007] A genomic portal system is described in accordance with another embodiment that includes an application server comprising an input manager that receives from a user a selection of one or more probe-set identifiers associated with one or more probe sets. The probe-set identifiers include one or more annotation terms. The system also includes a determiner that correlates the user-selected probe-set identifiers with one or more biological sequences, and a correlator that correlates the one or more biological sequences with a first set of probe sets. Another element of the genomic portal system is a network server comprising an output manager that identifies the first set of probe sets to the user.
- [0008] In accordance with another embodiment, a method is described for providing information related to one or more probe sets. Each probe-set has one or more identifiers. The method includes the acts of receiving from a user a selection of a first set of one or more of the probe-set identifiers, wherein each probe-set is capable of the identification of a biological molecule; correlating the first set of probe-set identifiers with a first set of one or more data; correlating each of the first set of data with a second set of one or more data; and providing the second set of data to the user.
- [0009] A system is described in accordance with yet another embodiment that includes an input manager that receives from a user a selection of one or more probe-set identifiers associated with one or more probe sets. The probe-set identifiers include

one or more biological sequences. Also included in the system is a database manager that periodically updates one or more local genomic databases, and a data generator that clusters the one or more biological sequences based at least in part on at least one of the local genomic databases. The system may also include a data processor that formats the clustered data for use in a graphical user interface. The graphical user interface may have one or more graphical elements representing protein family data, sequence alignment, or both. The sequence alignment graphical element may include a representation of a consensus sequence specific to a protein family. The system may also format the clustered data for storage and/or provide it to a user for display. The one or more biological sequences may include a protein sequence. In some implementations, the probe-set identifiers include one or more nucleotide sequences, and the system correlates the one or more nucleotide sequences with one or more protein sequences. In these implementations, the system clusters the one or more protein sequences based at least in part on at least one of the local genomic databases. The nucleotide sequence may include a gene, an EST, a consensus, or an SIF sequence, or any combination of one or more of them. The data generator may cluster the one or more protein sequences based, at least in part, on a learning algorithm, such as a Hidden Markov Model or a neural network as non-limiting examples.

- [0010] In accordance with other embodiments, a method is described for providing information related to one or more probe sets. The method includes the acts of receiving from a user a selection of one or more probe-set identifiers associated with one or more probe sets, periodically updating one or more local genomic databases; and clustering the one or more biological sequences based at least in part on at least one of the local genomic databases. The probe-set identifiers include one or more biological sequences.
- [0011] Also, a genomic portal system is described in accordance with some embodiments for providing information related to one or more probe sets. The system includes an application server and a network server. The application server receives from a user a selection of one or more probe-set identifiers associated with one or more probe sets, wherein the probe-set identifiers include one or more biological sequences, periodically updates one or more local genomic databases, clusters the one or more

biological sequences based at least in part on at least one of the local genomic databases, and formats the clustered data. The network server accepts the formatted cluster data from the data processor and provides the data to a user. Also described is a genomic portal system including an application server that receives from a user a plurality of probe-set identifiers associated with a plurality of probe sets. The probe-set identifiers are included in a batch file. The system periodically updates one or more local genomic databases and correlates each of the probe-set identifiers with data based on at least one of the local genomic databases. Also included in this system is a network server that provides the data to the user. It will be understood that the separation of the functions of the genomic portal system of these and other embodiments into an application server and a network server is illustrative only. In various implementations of these systems the functions performed by the two servers could be performed by a single server or other computing platform, distributed over more than two computer platforms, or otherwise distributed in accordance with various known computing techniques.

- [0012] In accordance with another embodiment, a method is described that includes the acts of receiving from a user a plurality of probe-set identifiers associated with a plurality of probe sets, wherein the probe-set identifiers are included in a batch file; periodically updating one or more local genomic databases; correlating each of the probe-set identifiers with data based on at least one of the local genomic databases; and providing the data to the user over a network.
- [0013] In a further embodiment, a system is described for providing information related to one or more probe sets. The system includes an input manager that receives one or more probe-set identifiers associated with one or more probe sets; a determiner that correlates the probe-set identifiers with a first set of data; and a correlator that correlates the first set of data with a second set of data. The probe-set identifiers include one or more of accession number, manufacturer-defined probe set identifier, biological sequence, or annotation term. The first set of data includes biological sequence or structure data, and the second set of data includes a probe-set identifier or protein information. The protein information may include biological process, molecular function, and/or cellular location information. The protein information may include information related to protein domain, sequence homology, complex

membership, pathway, biological system role, and/or interaction with other proteins or biological molecules. Also, a system is described in accordance with another embodiment including an input manager that receives one or more probe-set identifiers associated with one or more probe sets, and a manager that correlates the probe-set identifiers with a set of data. The probe-set identifiers include accession number, manufacturer-defined probe set identifier, biological sequence, and/or annotation term. The set of data includes protein information, which may include biological process, molecular function, or cellular location information. A method is described in accordance with another embodiment including the acts of receiving one or more probe-set identifiers associated with one or more probe sets and correlating the probe-set identifiers with a set of data. The probe-set identifiers include accession number, manufacturer-defined probe set identifier, biological sequence, and/or annotation term. The set of data includes protein information, which may include biological process, molecular function, or cellular location information.

[0014] In yet a further embodiment, a genomic portal system is described that includes an application server having an input manager that receives one or more probe-set identifiers associated with one or more probe sets, and a manager that correlates the probe-set identifiers with a set of data: Also included in the system is a network server that provides the data to the user. The probe-set identifiers include accession number, manufacturer-defined probe set identifier, biological sequence, and/or annotation term. The set of data includes protein information, such as biological process, molecular function, and/or cellular location information.

[0015] The above implementations are not necessarily inclusive or exclusive of each other and may be combined in any manner that is non-conflicting and otherwise possible, whether they be presented in association with a same, or a different, aspect or implementation. The description of one implementation is not intended to be limiting with respect to other implementations. Also, any one or more function, step, operation, or technique described elsewhere in this specification may, in alternative implementations, be combined with any one or more function, step, operation, or technique described in the summary. Thus, the above implementations are illustrative rather than limiting.

Brief Description of Drawings

- [0016] The above and further advantages will be more clearly appreciated from the following detailed description when taken in conjunction with the accompanying drawings. In the drawings, like reference numerals indicate like structures or method steps and the leftmost digit of a reference numeral indicates the number of the figure in which the referenced element first appears (for example, the element 180 appears first in Figure 1). In functional block diagrams, rectangles generally indicate functional elements, parallelograms generally indicate data, rectangles with curved sides generally indicate stored data, rectangles with a pair of double borders generally indicate predefined functional elements, and keystone shapes generally indicate manual operations. In method flow charts, rectangles generally indicate method steps and diamond shapes generally indicate decision elements. All of these conventions, however, are intended to be typical or illustrative, rather than limiting.
- [0017] Figure 1 is a functional block diagram of a probe-array analysis system including a scanner and a computer system on which may be executed computer applications suitable for providing probe-set identifiers and for receiving user selections of probe-set identifiers for processing;
- [0018] Figure 2 is a functional block diagram of one embodiment of probe-array analysis applications as illustratively stored for execution in system memory of the computer system of Figure 1;
- [0019] Figure 3 is a functional block diagram of a conventional system for obtaining genomic information over the Internet;
- [0020] Figure 4 is a functional block diagram of one embodiment of a genomic portal coupled over the Internet to remote databases and web pages and to clients including networks having user computer systems including that of Figure 1;
- [0021] Figure 5 is a functional block diagram of one embodiment of the genomic portal of Figure 4 including illustrative embodiments of a database server, portal application computer system, and portal-side Internet server;
- [0022] Figure 6 is a simplified graphical representation of one embodiment of computer application platforms for implementing the genomic portal of Figures 4 and 5 in

communication with clients such as those shown in Figure 4;

- [0023] Figure 7 is a flow chart of one embodiment of a method for providing a user with genomic product information related to gene expression, or differential expression, experimental results;
- [0024] Figure 8 is a functional block diagram of one embodiment of a user-service manager application as may be executed on the portal application computer system of Figure 5;
- [0025] Figure 9 is a simplified graphical representation of one embodiment of a gene or probe-set identifier to products and/or genomics database such as may be by the user-service manager of Figure 8;
- [0026] Figure 10 is a simplified graphical representation of one embodiment of a local genomic and/or product database such as may be accessed by the database server of figure 5;
- [0027] Figure 11 is a functional block diagram of one embodiment of a correlator such as may be included in the user-service manager application of Figure 8; and
- [0028] Figure 12 is a graphical representation of one embodiment of a graphical user interface suitable for providing genomic data to a user based on data correlated by the correlator of Figure 11.

Detailed Description

- [0029] Systems, methods, and computer products are now described with reference to an illustrative embodiment referred to as genomic portal 400. Portal 400 is shown in an Internet environment in Figure 4, and is illustrated in greater detail in Figures 5-12. In a typical implementation, portal 400 may be used to provide a user with information related to results from experiments with probe arrays. The experiments often involve the use of scanning equipment to detect hybridization of probe-target pairs, and the analysis of detected hybridization by various software applications, as now described in relation to Figures 1 and 2.

[0030]

Probe Arrays 103: Various techniques and technologies may be used for

synthesizing dense arrays of biological materials on or in a substrate or support. For example, Affymetrix® GeneChip® arrays are synthesized in accordance with techniques sometimes referred to as VLSIPS™ (Very Large Scale Immobilized Polymer Synthesis) technologies. Some aspects of VLSIPS™ and other microarray manufacturing technologies are described in U.S. Patents Nos. 5,424,186; 5,143,854; 5,445,934; 5,744,305; 5,831,070; 5,837,832; 6,022,963; 6,083,697; 6,291,183; 6,309,831; and 6,310,189, all of which are hereby incorporated by reference in their entireties for all purposes. The probes of these arrays in some implementations consist of nucleic acids that are synthesized by methods including the steps of activating regions of a substrate and then contacting the substrate with a selected monomer solution. As used herein, nucleic acids may include any polymer or oligomer of nucleosides or nucleotides (polynucleotides or oligonucleotides) that include pyrimidine and/or purine bases, preferably cytosine, thymine, and uracil, and adenine and guanine, respectively. Nucleic acids may include any deoxyribonucleotide, ribonucleotide, and/or peptide nucleic acid component, and/or any chemical variants thereof such as methylated, hydroxymethylated or glucosylated forms of these bases, and the like. The polymers or oligomers may be heterogeneous or homogeneous in composition, and may be isolated from naturally-occurring sources or may be artificially or synthetically produced. In addition, the nucleic acids may be DNA or RNA, or a mixture thereof, and may exist permanently or transitionally in single-stranded or double-stranded form, including homoduplex, heteroduplex, and hybrid states. Probes of other biological materials, such as peptides or polysaccharides as non-limiting examples, may also be formed. For more details regarding possible implementations, see U.S. Patent No. 6,156,501, which is hereby incorporated by reference herein in its entirety for all purposes.

[0031]

A system and method for efficiently synthesizing probe arrays using masks is described in U.S. Patent Application, Serial No. 09/824,931, filed April 3, 2001, that is hereby incorporated by reference herein in its entirety for all purposes. A system and method for a rapid and flexible microarray manufacturing and online ordering system is described in U.S. Provisional Patent Application, Serial No. 60/265,103, filed January 29, 2001, that also is hereby incorporated herein by reference in its entirety for all purposes. Systems and methods for optical photolithography without masks are

described in U.S. Patent No. 6,271,957 and in U.S. Patent Application No. 09/683,374 filed December 19, 2001, both of which are hereby incorporated by reference herein in their entireties for all purposes.

- [0032] The probes of synthesized probe arrays typically are used in conjunction with biological target molecules of interest, such as cells, proteins, genes or EST's, other DNA sequences, or other biological elements. More specifically, the biological molecule of interest may be a ligand, receptor, peptide, nucleic acid (oligonucleotide or polynucleotide of RNA or DNA), or any other of the biological molecules listed in U.S. Patent No. 5,445,934 (incorporated by reference above) at column 5, line 66 to column 7, line 51. For example, if transcripts of genes are the interest of an experiment, the target molecules would be the transcripts. Other examples include protein fragments, small molecules, etc. Target nucleic acid refers to a nucleic acid (often derived from a biological sample) of interest. Frequently, a target molecule is detected using one or more probes. As used herein, a probe is a molecule for detecting a target molecule. A probe may be any of the molecules in the same classes as the target referred to above. As non-limiting examples, a probe may refer to a nucleic acid, such as an oligonucleotide, capable of binding to a target nucleic acid of complementary sequence through one or more types of chemical bonds, usually through complementary base pairing, usually through hydrogen bond formation. As noted above, a probe may include natural (i.e. A, G, U, C, or T) or modified bases (7-deazaguanosine, inosine, etc.). In addition, the bases in probes may be joined by a linkage other than a phosphodiester bond, so long as the bond does not interfere with hybridization. Thus, probes may be peptide nucleic acids in which the constituent bases are joined by peptide bonds rather than phosphodiester linkages. Other examples of probes include antibodies used to detect peptides or other molecules, any ligands for detecting its binding partners. When referring to targets or probes as nucleic acids, it should be understood that these are illustrative embodiments that are not to limit the invention in any way.

- [0033] The samples or target molecules of interest (hereafter, simply targets) are processed so that, typically, they are spatially associated with certain probes in the probe array. For example, one or more tagged targets are distributed over the probe array. In accordance with some implementations, some targets hybridize with probes

and remain at the probe locations, while non-hybridized targets are washed away. These hybridized targets, with their tags or labels, are thus spatially associated with the probes. The hybridized probe and target may sometimes be referred to as a probe-target pair. Detection of these pairs can serve a variety of purposes, such as to determine whether a target nucleic acid has a nucleotide sequence identical to or different from a specific reference sequence. See, for example, U.S. Patent No. 5,837,832, referred to and incorporated above. Other uses include gene expression monitoring and evaluation (see, e.g., U.S. Patent No. 5,800,992 to Fodor, et al.; U.S. Patent No. 6,040,138 to Lockhart, et al.; and International App. No. PCT/US98/15151, published as WO99/05323, to Balaban, et al.), genotyping (U.S. Patent No. 5,856,092 to Dale, et al.), or other detection of nucleic acids. The '992, '138, and '092 patents, and publication WO99/05323, are incorporated by reference herein in their entireties for all purposes.

[0034]

Other techniques exist for depositing probes on a substrate or support. For example, "spotted arrays" are commercially fabricated, typically on microscope slides. These arrays consist of liquid spots containing biological material of potentially varying compositions and concentrations. For instance, a spot in the array may include a few strands of short oligonucleotides in a water solution, or it may include a high concentration of long strands of complex proteins. The Affymetrix ® 417 ™ Arrayer and 427 ™ Arrayer are devices that deposit densely packed arrays of biological materials on microscope slides in accordance with these techniques. Aspects of these, and other, spot arrayers are described in U.S. Patents Nos. 6,040,193 and 6,136,269; in U.S. Patent Application Serial No. 09/683,298; and in PCT Application No. PCT/US99/00730 (International Publication Number WO 99/36760), all of which are hereby incorporated by reference in their entireties for all purposes. Other techniques for generating spotted arrays also exist. For example, U.S. Patent No. 6,040,193 to Winkler, et al. is directed to processes for dispensing drops to generate spotted arrays. The '193 patent, and U.S. Patent No. 5,885,837 to Winkler, also describe the use of micro-channels or micro-grooves on a substrate, or on a block placed on a substrate, to synthesize arrays of biological materials. These patents further describe separating reactive regions of a substrate from each other by inert regions and spotting on the reactive regions. The '193 and '837 patents are hereby incorporated

5
6
7
8
9
10
11
12
13
14
15
16
17
18

by reference in their entireties. Another technique is based on ejecting jets of biological material to form a spotted array. Other implementations of the jetting technique may use devices such as syringes or piezo electric pumps to propel the biological material. It will be understood that the foregoing are non-limiting examples of techniques for synthesizing, depositing, or positioning biological material onto or within a substrate. For example, although a planar array surface is preferred in some implementations of the foregoing, a probe array may be fabricated on a surface of virtually any shape or even a multiplicity of surfaces. Arrays may comprise probes synthesized or deposited on beads, fibers such as fiber optics, glass or any other appropriate substrate, see U.S. Patent Nos. 6,361,947, 5,770,358, 5,789,162, 5,708,153 and 5,800,992, all of which are hereby incorporated in their entireties for all purposes. Arrays may be packaged in such a manner as to allow for diagnostics or other manipulation of in an all inclusive device, see for example, U.S. Pat. Nos. 5,856,174 and 5,922,591 incorporated in their entireties by reference for all purposes.

[0035] To ensure proper interpretation of the term "probe" as used herein, it is noted that contradictory conventions exist in the relevant literature. The word "probe" is used in some contexts to refer not to the biological material that is synthesized on a substrate or deposited on a slide, as described above, but to what has been referred to herein as the "target." To avoid confusion, the term "probe" is used herein to refer to probes such as those synthesized according to the VLSIPS™ technology; the biological materials deposited so as to create spotted arrays; and materials synthesized, deposited, or positioned to form arrays according to other current or future technologies. Thus, microarrays formed in accordance with any of these technologies may be referred to generally and collectively hereafter for convenience as "probe arrays." Moreover, the term "probe" is not limited to probes immobilized in array format. Rather, the functions and methods described herein may also be employed with respect to other parallel assay devices. For example, these functions and methods may be applied with respect to probe-set identifiers that identify probes immobilized on or in beads, optical fibers, or other substrates or media.

[0036] Probes typically are able to detect the expression of corresponding genes or EST's by detecting the presence or abundance of mRNA transcripts present in the target.

This detection may, in turn, be accomplished in some implementations by detecting labeled cRNA that is derived from cDNA derived from the mRNA in the target. In general, a group of probes, sometimes referred to as a probe set, contains subsequences in unique regions of the transcripts and does not correspond to a full gene sequence. Further details regarding the design and use of probes and probe sets are provided in U.S. Patent No. 6,188,783; in PCT Application Serial No. PCT/US 01/02316, filed January 24, 2001; and in U.S. Patent Applications Serial No. 09/721,042, filed on November 21, 2000, Serial No. 09/718,295, filed on November, 21, 2000, Serial No. 09/745,965, filed on December 21, 2000, and Serial No. 09/764,324, filed on January 16, 2001, all of which patents and patent applications are hereby incorporated herein by reference in their entireties for all purposes.

[0037]

Scanner 190: Figure 1 is a functional block diagram of a system that is suitable for, among other things, analyzing probe arrays that have been hybridized with labeled targets. Representative hybridized probe arrays 103 of Figure 1 may include probe arrays of any type, as noted above. Labeled targets in hybridized probe arrays 103 may be detected using various commercial devices, referred to for convenience hereafter as "scanners." An illustrative device is shown in Figure 1 as scanner 190. Scanners image the targets by detecting fluorescent or other emissions from the labels, or by detecting transmitted, reflected, or scattered radiation. These processes are generally and collectively referred to hereafter for convenience simply as involving the detection of "emissions." Various detection schemes are employed depending on the type of emissions and other factors. A typical scheme employs optical and other elements to provide excitation light and to selectively collect the emissions. Also generally included are various light-detector systems employing photodiodes, coupled devices, photomultiplier tubes, or similar devices to register the collected emissions. For example, a scanning system for use with a fluorescent label is described in U.S. Pat. No. 5,143,854, incorporated by reference above. Other scanners or scanning systems are described in U.S. Patent Nos. 5,578,832; 5,631,734; 5,834,758; 5,936,324; 5,981,956; 6,025,601; 6,141,096; 6,185,030; and 6,201,639; in PCT Application PCT/US99/ 06097 (published as WO99/47964); and in U.S. Patent Applications, Serial Nos. 09/682,837 filed October 23, 2001, 09/683,216 filed December 3, 2001, and 09/683,217 filed December 3, 2001, 09/683,219 filed

December 3, 2001, each of which is hereby incorporated by reference in its entirety for all purposes.

- [0038] Scanner 190 provides data representing the intensities (and possibly other characteristics, such as color) of the detected emissions, as well as the locations on the substrate where the emissions were detected. The data typically are stored in a memory device, such as system memory 120 of user computer 100, in the form of a data file. One type of data file, such as image data file 212 shown in Figure 2, typically includes intensity and location information corresponding to elemental sub-areas of the scanned substrate. The term "elemental" in this context means that the intensities, and/or other characteristics, of the emissions from this area each are represented by a single value. When displayed as an image for viewing or processing, elemental picture elements, or pixels, often represent this information. Thus, for example, a pixel may have a single value representing the intensity of the elemental sub-area of the substrate from which the emissions were scanned. The pixel may also have another value representing another characteristic, such as color. For instance, a scanned elemental sub-area in which high-intensity emissions were detected may be represented by a pixel having high luminance (hereafter, a "bright" pixel), and low-intensity emissions may be represented by a pixel of low luminance (a "dim" pixel). Alternatively, the chromatic value of a pixel may be made to represent the intensity, color, or other characteristic of the detected emissions. Thus, an area of high-intensity emission may be displayed as a red pixel and an area of low-intensity emission as a blue pixel. As another example, detected emissions of one wavelength at a particular sub-area of the substrate may be represented as a red pixel, and emissions of a second wavelength detected at another sub-area may be represented by an adjacent blue pixel. Many other display schemes are known. Two examples of image data are data files in the form *.dat or *.tif as generated respectively by Affymetrix[®] Microarray Suite based on images scanned from GeneChip[®] arrays, and by Affymetrix[®] Jaguar[™] software based on images scanned from spotted arrays.

- [0039] Probe-Array Analysis Applications 199: Generally, a human being may inspect a printed or displayed image constructed from the data in an image file and may identify those cells that are bright or dim, or are otherwise identified by a pixel characteristic (such as color). However, it frequently is desirable to provide this

information in an automated, quantifiable, and repeatable way that is compatible with various image processing and/or analysis techniques. For example, the information may be provided for processing by a computer application that associates the locations where hybridized targets were detected with known locations where probes of known identities were synthesized or deposited. Other methods include tagging individual synthesis or support substrates (such as beads) using chemical, biological, electro-magnetic transducers or transmitters, and other identifiers. Information such as the nucleotide or monomer sequence of target DNA or RNA may then be deduced. Techniques for making these deductions are described, for example, in U.S. Patent 5,733,729, which hereby is incorporated by reference in its entirety for all purposes, and in U.S. Patent No. 5,837,832, noted and incorporated above.

[0040]

A variety of computer software applications are commercially available for controlling scanners (and other instruments related to the hybridization process, such as hybridization chambers), and for acquiring and processing the image files provided by the scanners. Examples are the Jaguar™ application from Affymetrix, Inc., aspects of which are described in PCT Application PCT/US 01/26390 and in U.S. Patent Applications, Serial Nos. 09/681,819, 09/682,071, 09/682,074, and 09/682,076, and the Microarray Suite application from Affymetrix, aspects of which are described in U.S. Provisional Patent Applications, Serial Nos. 60/220,587, 60/220,645 and 60/312,906, all of which are hereby incorporated herein by reference in their entireties for all purposes. For example, image data in image data file 212 may be operated upon to generate intermediate results such as so-called cell intensity files (*.cel) and chip files (*.chp), generated by Microarray Suite or spot files (*.spt) generated by Jaguar™ software. For convenience, the terms "file" or "data structure" may be used herein to refer to the organization of data, or the data itself generated or used by executables 199A and executable counterparts of other applications. However, it will be understood that any of a variety of alternative techniques known in the relevant art for storing, conveying, and/or manipulating data may be employed, and that the terms "file" and "data structure" therefore are to be interpreted broadly. In the illustrative case in which image data file 212 is derived from a GeneChip® probe array, and in which Microarray Suite generates cell intensity file 216, file 216 may contain, for each probe scanned by scanner 190, a single value representative of

the intensities of pixels measured by scanner 190 for that probe. Thus, this value is a measure of the abundance of tagged cRNA's present in the target that hybridized to the corresponding probe. Many such cRNA's may be present in each probe, as a probe on a GeneChip® probe array may include, for example, millions of oligonucleotides designed to detect the cRNA's. The resulting data stored in the chip file may include degrees of hybridization, absolute and/or differential (over two or more experiments) expression, genotype comparisons, detection of polymorphisms and mutations, and other analytical results. In another example, in which executables 199A includes image data from a spotted probe array, the resulting spot file includes the intensities of labeled targets that hybridized to probes in the array. Further details regarding cell files, chip files, and spot files are provided in U.S. Provisional Patent Application Nos. 60/220,645, 60/220,587, and 60/226,999, incorporated by reference above.

[0041]

In the present example, in which executables 199A include Affymetrix® Microarray Suite, the chip file is derived from analysis of the cell file combined in some cases with information derived from library files (not shown) that specify details regarding the sequences and locations of probes and controls. Laboratory or experimental data may also be provided to the software for inclusion in the chip file. For example, an experimenter and/or automated data input devices or programs (not shown) may provide data related to the design or conduct of experiments. As a non-limiting example related to the processing of an Affymetrix® GeneChip® probe array, the experimenter may specify an Affymetrix catalogue or custom chip type (e.g., Human Genome U95Av2 chip) either by selecting from a predetermined list presented by Microarray Suite or by scanning a bar code related to a chip to read its type. Microarray Suite may associate the chip type with various scanning parameters stored in data tables including the area of the chip that is to be scanned, the location of chrome borders on the chip used for auto-focusing, the wavelength or intensity of laser light to be used in reading the chip, and so on. Other experimental or laboratory data may include, for example, the name of the experimenter, the dates on which various experiments were conducted, the equipment used, the types of fluorescent dyes used as labels, protocols followed, and numerous other attributes of experiments. As noted, executables 199A may apply some of this data in the generation of intermediate results. For example, information about the dyes may be

incorporated into determinations of relative expression. Other data, such as the name of the experimenter, may be processed by executables 199A or may simply be preserved and stored in files or other data structures. Any of these data may be provided, for example over a network, to a laboratory information management server computer, such as user database server 412 of Figure 4, configured to manage information from large numbers of experiments. Data analysis program 210 may also generate various types of plots, graphs, tables, and other tabular and/or graphical representations of analytical data such as contained in file 215. As will be appreciated by those skilled in the relevant art, the preceding and following descriptions of files generated by executables 199A are exemplary only, and the data described, and other data, may be processed, combined, arranged, and/or presented in many other ways.

[0042] The processed image files produced by these applications often are further processed to extract additional data. In particular, data-mining software applications often are used for supplemental identification and analysis of biologically interesting patterns or degrees of hybridization of probe sets. An example of a software application of this type is the Affymetrix ® Data Mining Tool, described in U.S. Provisional Patent Applications, Serial Nos. 60/274,986 and 60/312,256, both of which are hereby incorporated herein by reference in their entireties for all purposes. Software applications also are available for storing and managing the enormous amounts of data that often are generated by probe-array experiments and by the image-processing and data-mining software noted above. An example of these data-management software applications is the Affymetrix ® Laboratory Information Management System (LIMS), aspects of which are described in U.S. Patent Application No. 09/682,098 and in U.S. Provisional Patent Applications, Serial Nos. 60/220,587 and 60/220,645, all of which are hereby incorporated by reference herein in their entireties for all purposes. In addition, various proprietary databases accessed by database management software, such as the Affymetrix ® EASI (Expression Analysis Sequence Information) database and database software, provide researchers with associations between probe sets and gene or EST identifiers.

[0043] For convenience of reference, these types of computer software applications (i.e., for acquiring and processing image files, data mining, data management, and various database and other applications related to probe-array analysis) are generally and

collectively represented in Figure 1 as probe-array analysis applications 199. Figure 2 is a functional block diagram of probe-array analysis applications 199 as illustratively stored for execution (as executable code 199A corresponding to applications 199) in system memory 120 of user computer 100 of Figure 1.

[0044] As will be appreciated by those skilled in the relevant art, it is not necessary that applications 199 be stored on and/or executed from computer 100; rather, some or all of applications 199 may be stored on and/or executed from an applications server or other computer platform to which computer 100 is connected in a network. For example, it may be particularly advantageous for applications involving the manipulation of large databases, such as Affymetrix ® LIMS or Affymetrix ® Data Mining Tool (DMT), to be executed from a database server such as user database server 412 of Figure 4. Alternatively, LIMS, DMT, and/or other applications may be executed from computer 100, but some or all of the databases upon which those applications operate may be stored for common access on server 412 (perhaps together with a database management program, such as the Oracle ® 8.0.5 database management system from Oracle Corporation). Such networked arrangements may be implemented in accordance with known techniques using commercially available hardware and software, such as those available for implementing a local-area network or wide-area network. A local network is represented in Figure 4 by the connection of user computer 100 to user database server 412 (and to user-side Internet client 410, which may be the same computer) via network cable 480. Similarly, scanner 190 (or multiple scanners) may be made available to a network of users over cable 480 both for purposes of controlling scanner 190 and for receiving data input from it.

[0045] In some implementations, it may be convenient for user 101 to group probe-set identifiers 222 for batch transfer of information or to otherwise analyze or process groups of probe sets together. For example, as described below, user 101 may wish to obtain annotation information via portal 400 related to one or more probe sets identified by their respective probe set identifiers. Rather than obtaining this information serially, user 101 may group probe sets together for batch processing. Various known techniques may be employed for associating probe set identifiers, or data related to those identifiers, together. For instance, user 101 may generate a tab delimited *.txt file including a list of probe set identifiers for batch processing. This

PCT/US2008/063559

file or another file or data structure for providing a batch of data (hereafter referred to for convenience simply as a "batch file"), may be any kind of list, text, data structure, or other collection of data in any format. The batch file may also specify what kind of information user 101 wishes to obtain with respect to all, or any combination of, the identified probe sets. In some implementations, user 101 may specify a name or other user-specified identifier to represent the group of probe-set identifiers specified in the text file or otherwise specified by user 101. This user-specified identifier may be stored by one of executables 199A, or by elements of portal 400 described below, so that user 101 may employ it in future operations rather than providing the associated probe-set identifiers in a text file or other format. Thus, for example, user 101 may formulate one or more queries associated with a particular user-specified identifier, resulting in a batch transfer of information from portal 400 to user 101 related to the probe-set identifiers that user 101 has associated with the user-specified identifier. Alternatively, user 101 may initiate a batch transfer by providing the text file of probe-set identifiers. In any of these cases, user 101 may formulate queries to obtain, in a single batch operation, probe set records, lists of probe sets sorted into functional groups, protein domain information, sequence homology information, metabolic pathway information, BLAST similarity searches, array content information, and any other information available via portal 400. Similarly, user 101 may provide information, such as laboratory or experimental information, related to a number of probe sets by a batch operation rather than serial ones. The probe sets may be grouped by experiments, by similarity of probe sets (e.g., probe sets representing genes having similar annotations, such as related to transcription regulation), or any other type of grouping. For example, user 101 may assign a user-specified identifier (e.g., "experiments of January 1") to a series of experiments and submit probe-set identifiers in user-selected categories (e.g., identifying probe sets that were up-regulated by a specified amount) and provide the experimental information to portal 400 for data storage and/or analysis.

[0046]

User Computer 100: User computer 100, shown in Figure 1, may be a computing device specially designed and configured to support and execute some or all of the functions of probe array applications 199. Computer 100 also may be any of a variety of types of general-purpose computers such as a personal computer, network server,

workstation, or other computer platform now or later developed. Computer 100 typically includes known components such as a processor 105, an operating system 110, a graphical user interface (GUI) controller 115, a system memory 120, memory storage devices 125, and input-output controllers 130. It will be understood by those skilled in the relevant art that there are many possible configurations of the components of computer 100 and that some components that may typically be included in computer 100 are not shown, such as cache memory, a data backup unit, and many other devices. Processor 105 may be a commercially available processor such as a Pentium ® processor made by Intel Corporation, a SPARC ® processor made by Sun Microsystems, or it may be one of other processors that are or will become available. Processor 105 executes operating system 110, which may be, for example, a Windows ® -type operating system (such as Windows NT ® 4.0 with SP6a) from the Microsoft Corporation; a Unix ® or Linux-type operating system available from many vendors; another or a future operating system; or some combination thereof. Operating system 110 interfaces with firmware and hardware in a well-known manner, and facilitates processor 105 in coordinating and executing the functions of various computer programs that may be written in a variety of programming languages. Operating system 110, typically in cooperation with processor 105, coordinates and executes functions of the other components of computer 100. Operating system 110 also provides scheduling, input-output control, file and data management, memory management, and communication control and related services, all in accordance with known techniques.

[0047]

System memory 120 may be any of a variety of known or future memory storage devices. Examples include any commonly available random access memory (RAM), magnetic medium such as a resident hard disk or tape, an optical medium such as a read and write compact disc, or other memory storage device. Memory storage device 125 may be any of a variety of known or future devices, including a compact disk drive, a tape drive, a removable hard disk drive, or a diskette drive. Such types of memory storage device 125 typically read from, and/or write to, a program storage medium (not shown) such as, respectively, a compact disk, magnetic tape, removable hard disk, or floppy diskette. Any of these program storage media, or others now in use or that may later be developed, may be considered a computer program product.

As will be appreciated, these program storage media typically store a computer software program and/or data. Computer software programs, also called computer control logic, typically are stored in system memory and/or the program storage device used in conjunction with memory storage device 125.

[0048] In some embodiments, a computer program product is described comprising a computer usable medium having control logic (computer software program, including program code) stored therein. The control logic, when executed by processor 105, causes processor 105 to perform functions described herein. In other embodiments, some functions are implemented primarily in hardware using, for example, a hardware state machine. Implementation of the hardware state machine so as to perform the functions described herein will be apparent to those skilled in the relevant arts.

[0049] Input-output controllers 130 could include any of a variety of known devices for accepting and processing information from a user, whether a human or a machine, whether local or remote. Such devices include, for example, modem cards, network interface cards, sound cards, or other types of controllers for any of a variety of known input devices 102. Output controllers of input-output controllers 130 could include controllers for any of a variety of known display devices 180 for presenting information to a user, whether a human or a machine, whether local or remote. If one of display devices 180 provides visual information, this information typically may be logically and/or physically organized as an array of picture elements, sometimes referred to as pixels. Graphical user interface (GUI) controller 115 may comprise any of a variety of known or future software programs for providing graphical input and output interfaces between computer 100 and user 101, and for processing user inputs. In the illustrated embodiment, the functional elements of computer 100 communicate with each other via system bus 104. Some of these communications may be accomplished in alternative embodiments using network or other types of remote communications.

[0050] As will be evident to those skilled in the relevant art, applications 199, if implemented in software, may be loaded into system memory 120 and/or memory storage device 125 through one of input devices 102. All or portions of applications 199 may also reside in a read-only memory or similar device of memory storage

device 125, such devices not requiring that applications 199 first be loaded through input devices 102. It will be understood by those skilled in the relevant art that applications 199, or portions of it, may be loaded by processor 105 in a known manner into system memory 120, or cache memory (not shown), or both, as advantageous for execution.

- [0051] Conventional Techniques for Obtaining Genomic Data: A number of conventional approaches for obtaining genomic data over the Internet are available, some of which are described in the book edited by Ouellette and Bzevanis, incorporated by reference above. Figure 3 is a functional block diagram representing one simplified example. As shown in Figure 3, user 101 may consult any of a number of public or other sources to obtain accession numbers 224'. As represented by manual operation 312, user 101 initiates request 312 by accessing through any web browser the Internet web site of the National Center for Biotechnology Information (NCBI) of the National Library of Medicine and the National Institutes of Health (as of January 2001, accessible at the Internet URL <http://www.ncbi.nlm.nih.gov/>). In particular, user 101 may access the Entrez search and retrieval system that provides information from various databases at NCBI . These databases provide information regarding nucleotide sequences, protein sequences, macromolecular structures, whole genomes, and publication data related thereto. It is illustratively assumed that user 101 accesses in this manner NCBI Entrez nucleotide database 314 and receives information including gene or EST sequences 316. Particularly if accession numbers 224' represents a large number (e.g., one hundred) of EST's or genes of interest, as may easily be the case following analysis of probe array experiments, the tasks thus far described may take significant time, perhaps hours.
- [0052] User 101 typically copies sequence information from sequences 316 and pastes this information into an HTML document accessible through NCBI's BLAST web pages 324 (as of April 2002, accessible at <http://www.ncbi.nlm.nih.gov/BLAST/>). This operation, which also may be time consuming and tedious if many sequences are involved, is represented by user-initiated batch BLAST request 322 of Figure 3. BLAST is an acronym for Basic Local Alignment Search Tool, and, as is well known in the art, consists of similarity search programs that interrogate sequence databases for both protein and DNA using heuristic algorithms to seek local alignments. For example,

user 101 may conduct a BLAST search using the "blastn" nucleotide sequence database. Results of this batch BLAST search, represented by similar nucleotide and/or protein sequence data 326, on occasion may not be available to user 101 for many minutes or even hours. User 101 may then initiate comparisons and evaluations 332, which may be conducted manually or using various software tools. User 101 may subsequently issue report 334 interpreting the findings of the searches and positing strategies and requirements for follow-on experiments.

- [0053] Inputs to Genomic Portal 400 from User 101: Figure 4 is a functional block diagram showing an illustrative configuration by which user 101 may connect with genomic web portal 400. It will be understood that Figure 4 is simplified and is illustratively only, and that many implementations and variations of the network and Internet connections shown in Figure 4 will be evident to those of ordinary skill in the relevant art.
- [0054] User 101 employs user computer 100 and analysis applications 199 as noted above, including generating and/or accessing some or all of files 212-217. As shown in Figure 4, files 212-217 are maintained in this example on user database server 412 to which user computer 100 is coupled via network cable 480. Computers 100', 100'', and computers of other users in a local or wide-area network including an Intranet, the Internet, or any other network may also be coupled to server 412 via cable 480. It will be understood that cable 400 is merely representative of any type of network connectivity, which may involve cables, transmitters, relay stations, network servers, and many other components not shown but evident to those of ordinary skill in the relevant art. Via user computer 100, user 101 may operate a web browser served by user-side Internet client 410 to communicate via Internet 499 with portal 400. Portal 400 may similarly be in communication over Internet 499 with other users and/or networks of users, as indicated by Internet clients 410' and 410''.

- [0055] As previously noted, the information provided by user 101 to portal 400 typically includes one or more "probe-set identifiers." These probe-set identifiers typically come to the attention of user 101 as a result of experiments conducted on probe arrays. For example, user 101 may select probe-set identifiers that identify microarray probe sets capable of enabling detection of the expression of mRNA transcripts from

corresponding genes or EST's of particular interest. As is well known in the relevant art, an EST is a fragment of a gene sequence that may not be fully characterized, whereas a gene sequence generally is complete and fully characterized. The word "gene" is used generally herein to refer both to full size genes of known sequence and to computationally predicted genes. In some implementations, the specific sequences detected by the arrays that represent these genes or EST's may be referred to as, "sequence information fragments (SIF's)" and may be recorded in a "SIF file," as noted above with respect to the operations of LIMS 225. In particular implementations, a SIF is a portion of a consensus sequence that has been deemed to best represent the mRNA transcript from a given gene or EST. The consensus sequence may have been derived by comparing and clustering EST's, and possibly also by comparing the EST's to genomic sequence information. A SIF is a portion of the consensus sequence for which probes on the array are specifically designed. With respect to the operations of web portal 400, it is assumed with respect to some implementations that some microarray probe sets may be designed to detect the expression of genes based upon sequences of EST's.

[0056]

As was described above, the term "probe set" refers in some implementations to one or more probes from an array of probes on a microarray. For example, in an Affymetrix ® GeneChip ® probe array, in which probes are synthesized on a substrate, a probe set may consist of 30 or 40 probes, half of which typically are controls. These probes collectively, or in various combinations of some or all of them, are deemed to be indicative of the expression of a gene or EST. In a spotted probe array, one or more spots may similarly constitute a "probe set." The term "probe-set identifiers" is used broadly herein in that a number of types of such identifiers are possible and are intended to be included within the meaning of this term. One type of probe-set identifier is a name, number, or other symbol that is assigned for the purpose of identifying a probe set. This name, number, or symbol may be arbitrarily assigned to the probe set by, for example, the manufacturer of the probe array. A user may select this type of probe-set identifier by, for example, highlighting or typing the name. Another type of probe-set identifier as intended herein is a graphical representation of a probe set. For example, dots may be displayed on a scatter plot or other diagram wherein each dot represents a probe set. Typically, the dot's placement on the plot

represents the intensity of the signal from hybridized, tagged, targets (as described in greater detail below) in one or more experiments. In these cases, a user may select a probe-set identifier by clicking on, drawing a loop around, or otherwise selecting one or more of the dots. In another example, user 101 may select a probe-set identifier by selecting a row or column in a table or spreadsheet that correlates probe sets with accession numbers and other genomic information.

[0057] Yet another type of probe-set identifier, as that term is used herein, includes a nucleotide or amino acid sequence. For example, it is illustratively assumed that a particular SIF is a unique sequence of 500 bases that is a portion of a consensus sequence or exemplar sequence gleaned from EST and/or genomic sequence information. It further is assumed that one or more probe sets are designed to represent the SIF. A user who specifies all or part of the 500-base sequence thus may be considered to have specified all or some of the corresponding probe sets.

[0058] In yet another example, a user may specify an SIF, gene, protein, or EST sequence for which there are no corresponding probe sets. The user requests to have a corresponding probe set produced for the specified sequence. User-service manager 522 (described below) assigns an identifier for the new probe set and this identifier, together with the sequence or sequences from which the probes are to be designed, are stored by database manager 512 in one or more databases. Manager 522 designs probe sets for the corresponding SIF, gene, or EST and correlates the probe sets with the new probe set identifiers. Further details regarding the processing and implementation of custom probe designs are provided in U.S. Provisional Patent Application, Serial No. 60/301,298, incorporated by reference above.

[0059] As a further example with respect to a particular implementation, a user may specify a portion of the 500-base sequence noted above, which may be unique to that SIF, or, alternatively, may also identify another SIF, EST, cluster of EST's, consensus sequence, and/or gene or protein. The user thus specifies a probe-set identifier for one or more genes or EST's. In another variation, it is illustratively assumed that a particular SIF is a portion of a particular consensus sequence. It is further assumed that a user specifies a portion of the consensus sequence that is not included in the SIF but that is unique to the consensus sequence or the gene or EST's the consensus

sequence is intended to represent. In that case, the sequence specified by the user is a probe-set identifier that identifies the probe set corresponding to the SIF, even though the user-specified sequence is not included in the SIF. Parallel cases are possible with respect to user specifications of partial sequences of EST's and genes or EST's, as those skilled in the relevant art will now appreciate.

- [0060] A further example of a probe-set identifier is an accession number of a gene or EST. Gene and EST accession numbers are publicly available. A probe set may therefore be identified by the accession number or numbers of one or more EST's and/or genes corresponding to the probe set. The correspondence between a probe set and EST's or genes may be maintained in a suitable database, such as that accessed by database application 230 or local library databases 516, from which the correspondence may be provided to the user. Similarly, gene fragments or sequences other than EST's may be mapped (e.g., by reference to a suitable database) to corresponding genes or EST's for the purpose of using their publicly available accession numbers as probe-set identifiers. For example, a user may be interested in product or genomic information related to a particular SIF that is derived from EST-1 and EST-2. The user may be provided with the correspondence between that SIF (or part or all of the sequence of the SIF) and EST-1 or EST-2, or both. To obtain product or genomic data related to the SIF, or a partial sequence of it, the user may select the accession numbers of EST-1, EST-2, or both.
- [0061] Additional examples of probe-set identifiers include one or more terms that may be associated with the annotation of one or more gene or EST sequences, where the gene or EST sequences may be associated with one or more probe sets. For convenience, such terms may hereafter be referred to as "annotation terms" and will be understood to potentially include, in various implementations, one or more words, graphical elements, characters, or other representational forms that provide information that typically is biologically relevant to or related to the gene or EST sequence. Associations between the probe-set identifier terms and gene or EST sequences may be stored in a database such as Probe-set ID to sequence database 511, local genomic database 518, or they may be transferred from remote databases 402. Examples of such terms associated with annotations include those of molecular function (e.g. transcription initiation), cellular location (e.g. nuclear membrane),

biological process (e.g. immune response), tissue type (e.g. kidney), or other annotation terms known to those in the relevant art.

- [0062] To provide a further specific example, user 101 may input the illustrative annotation term "tumor suppression." A large number of genes or EST's are known to be involved with this biological process. For example, a gene known as p53 is involved with tumor suppression, and this information is stored in one or more of the databases accessible from database server 410. Portal 400 provides to user 101 a list of probe-set identifiers that includes the one or more probe-set identifiers associated with gene p53. The list of probe-set identifiers may be provided to the user in one of numerous possible formats. For example, the format may include a table comprising all the probe sets associated with all the genes or EST's associated with "tumor suppression." Alternatively, the format may separate the probe sets related to each gene or EST into its own table.
- [0063] Genomic web portal 400: Genomic web portal 400 provides to user 101 data related to one or more genes or EST's. Typically, each gene or EST has at least one corresponding probe set that is identified by a probe-set identifier that, as just noted, may be a number, name, accession number, symbol, graphical representation (e.g., dot or highlighted tabular entry), or nucleotide sequence, as illustrative and non-limiting examples. The corresponding probe sets are capable of enabling detection of the expression of their corresponding gene. In response to a user selection of one or more probe-set identifiers, portal 400 provides user 101 with genomic information and/or information regarding biological products. This information may be helpful to user 101 in analyzing the results of experiments and in designing or implementing follow-up experiments.
- [0064]

Figure 5 is a functional block diagram of one of many possible embodiments of portal 400. In this example, portal 400 has hardware components including three computer platforms: database server 510, Internet server 530, and application server 520. Various functional elements of portal 400, such as database manager 512, input and output managers 532 and 534, and user-service manager 522, carry out their operations on these computer platforms. That is, in a typical implementation, the functions of managers 512, 532, 534, and 522 are carried out by the execution of

software applications on and across the computer platforms represented by servers 510, 530, and 520. Portal 400 is described first with respect to its computer platforms, and then with respect to its functional elements.

[0065] Each of servers 510, 520 and 530 may be any type of known computer platform or a type to be developed in the future, although they typically will be of a class of computer commonly referred to as servers. However, they may also be a main frame computer, a work station, or other computer type. They may be connected via any known or future type of cabling or other communication system, either networked or otherwise. They may be co-located or they may be physically separated. Various operating systems may be employed on any of the computer platforms, possibly depending on the type and/or make of computer platform chosen. Appropriate operating systems include Windows NT®, Sun Solaris, Linux, OS/400, Compaq Tru64 Unix, SGI IRIX, Siemens Reliant Unix, and others.

[0066] There may be significant advantages to carrying out the functions of portal 400 on multiple computer platforms in this manner, such as lower costs of deployment, database switching, or changes to enterprise applications, and/or more effective firewalls. Other configurations, however, are possible. For example, as is well known to those of ordinary skill in the relevant art, so-called two-tier or N-tier architectures are possible rather than the three-tier server-side component architecture represented by Figure See, for example, E. Roman, Mastering Enterprise JavaBeans™ and the Java™ 2 Platform (John Wiley & Sons, Inc., NY, 1999) and J. Schneider and R. Arora, Using Enterprise Java™ (Que Corporation, Indianapolis, 1997), both of which are hereby incorporated by reference in their entireties for all purposes.

[0067] It will be understood that many hardware and associated software or firmware components that may be implemented in a server-side architecture for Internet commerce are not shown in Figure 5. Components to implement one or more firewalls to protect data and applications, uninterruptable power supplies, LAN switches, web-server routing software, and many other components are not shown. Similarly, a variety of computer components customarily included in server-class computing platforms, as well as other types of computers, will be understood to be included but are not shown. These components include, for example, processors, memory units,

input/output devices, buses, and other components noted above with respect to user computer 103. Those of ordinary skill in the art will readily appreciate how these and other conventional components may be implemented.

- [0068] The functional elements of portal 400 also may be implemented in accordance with a variety of software facilitators and platforms (although it is not precluded that some or all of the functions of portal 400 may also be implemented in hardware or firmware). Among the various commercial products available for implementing e-commerce web portals are BEA WebLogic from BEA Systems, which is a so-called "middleware" application. This and other middleware applications are sometimes referred to as "application servers," but are not to be confused with application server 520, which is a computer. The function of these middleware applications generally is to assist other software components (such as managers 512, 522, or 532) to share resources and coordinate activities. The goals include making it easier to write, maintain, and change the software components; to avoid data bottlenecks; and prevent or recover from system failures. Thus, these middleware applications may provide load-balancing, fail-over, and fault tolerance, all of which features will be appreciated by those of ordinary skill in the relevant art.
- [0069] Other development products, such as the Java™ 2 platform from Sun Microsystems, Inc. may be employed in portal 400 to provide suites of applications programming interfaces (API's) that, among other things, enhance the implementation of scalable and secure components. The platform known as J2EE (Java™ 2, Enterprise Edition), is configured for use with Enterprise JavaBeans™, both from Sun Microsystems. Enterprise JavaBeans™ generally facilitates the construction of server-side components using distributed object applications written in the Java™ language. Thus, in one implementation, the functional elements of portal 400 may be written in Java and implemented using J2EE and Enterprise JavaBeans™. Various other software development approaches or architectures may be used to implement the functional elements of portal 400 and their interconnection, as will be appreciated by those of ordinary skill in the art.

- [0070] One implementation of these platforms and components is shown in Figure 6. Figure 6 is a simplified graphical representation of illustrative interactions between

PCT/US2003/035333

user-side internet client 410 on the user side and input and output managers 532 and 534 of Internet server 530 on the portal side, as well as communications among the three tiers (servers 510, 520, and 530) of portal 400. Browser 605 on client 410 sends and receives HTML documents 620 to and from server 530. HTML document 625 includes applet 627. Browser 605, running on user computer 103, provides a run-time container for applet 627. Functions of managers 532 and 534 on server 530, such as the performance of GUI operations, may be implemented by servlet and/or JSP 640 operating with a Java™ platform. A servlet engine executing on server 530 provides a runtime container for servlet 640. JSP (Java Server Pages) from Sun Microsystems, Inc. is a script-like environment for GUI operations; an alternative is ASP (Active Server Pages) from the Microsoft Corporation. App server 650 is the middleware product referred to above, and executes on application server 520. EJB (Enterprise JavaBeans™) is a standard that defines an architecture for enterprise beans, which are application components. CORBA (Common Object Request Broker Architecture) similarly is a standard for distributed object systems, i.e., the CORBA standards are implemented by CORBA-compliant products such as Java™ IDL. An example of an EJB-compliant product is WebLogic, referred to above. Further details of the implementation of standards, platforms, components, and other elements for an Internet portal and its communications with clients, are well known to those skilled in the relevant art.

[0071]

As noted, one of the functional elements of portal 400 is input manager 532. Manager 532 receives a set, i.e., one or more, of probe-set identifiers from user 101 over Internet 499. Manager 532 processes and forwards this information to user-service manager 522. These functions are performed in accordance with known techniques common to the operation of Internet servers, also commonly referred to in similar contexts as presentation servers. Another of the functional elements of portal 400 is output manager 534. Manager 534 provides information assembled by user-service manager 522 to user 101 over Internet 499, also in accordance with those known techniques, aspects of which were described above in relation to Figure 6. The information assembled by manager 522 is represented in Figure 5 as data 524, labeled "integrated genomic and/or product web pages responsive to user request." The data is integrated in the sense, among other things, that it is based, at least in

part, on the specification by user 101 of probe-set identifiers and thus has common relationships to the genes and/or EST's corresponding to those identifiers. The presentation by manager 534 of data 524 may be implemented in accordance with a variety of known techniques. As some examples, data 524 may include HTML or XML documents, email or other files, or data in other forms. The data may include Internet URL addresses so that user 101 may retrieve additional HTML, XML, or other documents or data from remote sources.

[0072] Portal 400 further includes database manager 512. In the illustrated embodiment, database manager 512 coordinates the storage, maintenance, supplementation, and all other transactions from or to any of local databases 511, 513, 514, 516, and 518. Manager 512 may undertake these functions in cooperation with appropriate database applications such as the Oracle ® 8.0.5 database management system.

[0073] In some implementations, manager 512 periodically updates local genomic database 518. The data updated in database 518 includes data related to genes or EST's that correspond with one or more probe sets. The probe sets may be those used or designed for use on any microarray product, and/or that are expected or calculated to be used in microarray products of any manufacturer or researcher. For example, the probe sets may include all probe sets synthesized on the line of stocked GeneChip ® probe arrays from Affymetrix, Inc., including its Arabidopsis Genome Array, CYP450 Array, Drosophila Genome Array, E. coli Genome Array, GenFlex ™ Tag Array, HIV PRT Plus Array, HuGeneFL Array, Human Genome U95 Set, Human Genome U133 Set, HuSNP Probe Array, Murine Genome U74 Set, P53 Probe Array, Rat Genome U34 Set, Rat Neurobiology U34 Set, Rat Toxicology U34 Array, or Yeast Genome S98 Array. The probe sets may also include those synthesized on custom arrays for user 101 or others. However, the data updated in database 518 need not be so limited. Rather, it may relate to any number of genes or EST's. Types of data that may be stored in database 518 are described below in relation to the operations of manager 522 in directing the periodic collection of this data from remote sources providing the locally maintained data in database 518 to users.

[0074] Database 516 includes data of a type referred to above in relation to database application 230, i.e., data that associates probe sets with their corresponding gene or

EST and their identifiers. Database 516 may also include SIF's, and other library data. User-service manager 522 may provide database manager 512 from time to time with update information regarding library and other data. In some cases, this update information will be provided by the owners or managers of proprietary information, although this information may also be made available publicly, as on a web site, for uploading.

[0075] Information for storage by manager 512 in local products database 514 may similarly be provided by vendors, distributors, or agents, or obtained from public sources such as web sites. A wide variety of product-related information may be included in database 514, examples of which include availability, pricing, composition, suitability, or ordering data. The information may relate to a wide variety of products, including any type of biological device or substance, or any type of reagent that may be used with a biological device or substance. To provide just a few examples, the device, substance, or reagent may be an oligonucleotide, probe array, clone, antibody, or protein. The data stored in database 514 may also include links, such as Internet URL addresses, to remote sites where product data is available, such as vendors' web sites.

[0076] Database 511 includes information relating probe-set identifiers to the sequences of the probes. This information may be provided by the manufacturer of the probes, the researchers who devise probes for spotted arrays or other custom arrays, or others. Moreover, the application of portal 400 is not limited to probes arranged in arrays. As noted, probes may be immobilized on or in beads, optical fibers, or other substrates or media. Thus, database 511 may also include information regarding the sequences of these probes.

[0077] Database 519 includes information about users and their accounts for doing business with or through portal 400. Any of a variety of account information, such as current orders, past orders, and so on, may be obtained from users, all as will be readily apparent to those of ordinary skill in the art. Also, information related to users may be developed by recording and/or analyzing the interactions of users with portal 400, in accordance with known techniques used in e-commerce. For example, user-service manager 522 may take note of users' areas of genomic interest, their

purchase or product-inquiry activities, the frequency of their accessing of various services, and so on, and provide this information to database manager 512 for storage or update in database 519.

- [0078] Another functional element of portal 400 is user-service manager 522. Among other functions, manager 522 may periodically cause database manager 512 to update local genomic database 518 from various sources, such as remote databases 402. For example, according to any chronological schedule (e.g., daily, weekly, etc.), manager 522 may, in accordance with known techniques, initiate searches of remote databases 402 by formulating appropriate queries, addressed to the URL's of the various databases 402, or by other conventional techniques for conducting data searches and/or retrieving data or documents over the Internet. These search queries and corresponding addresses may be provided in a known manner to output manager 534 for presentation to databases 402. Input manager 532 receives replies to the queries and provides them to manager 522, which then provides them to database manager 512 for updating of database 518, all in accordance with any of a variety of known techniques for managing information flow to, from, and within an Internet site.
- [0079] Portal application manager 526 manages the administrative aspects of portal 400, possibly with the assistance of a middleware product such as an applications server product. One of these administrative tasks may be the issuance of periodic instructions to manager 522 to initiate the periodic updating of database 518 just described. Alternatively, manager 522 may self-initiate this task. It is not required that all data in database 518 be updated according to the same periodic schedule. Rather, it may be typical for different types of data and/or data from different sources to be updated according to different schedules. Moreover, these schedules may be changed, and need not be according to a consistent schedule. That is, updating for particular data may occur after a day, then again after 2 days, then at a different period that may continue to vary. Numerous factors may influence the determination by manager 526 or manager 522 to maintain or vary these periods, such as the response time from various remote databases 402, the value and/or timeliness of the information in those databases, cost considerations related to accessing or licensing the databases, the quantity of information that must be accessed, and so on.

[0080] In some implementations, manager 522 constructs from data in local genomic database 518 a set of data related to genes or EST's corresponding to the set of probe-set identifiers selected by user 101. The user selection may be forwarded to manager 522 by input manager 532 in accordance with known techniques. Manager 522, also in accordance with known techniques, obtains the data from database 518 by forming appropriate queries, such as in one of the varieties of SQL language, based on the user selection. Manager 522 then forwards the queries to database manager 512 for execution against database 518.

[0081] As noted, various types of data may be accessed from remote databases 402 and maintained in local genomic database 518. Examples are illustrated in figure 10 that include sequence data 1010, exonic structure or location data 1015, splice-variants data 1020, marker structure or location data 1025, polymorphism data 1030, homology data 1035, protein-family classification data 1040, pathway data 1045, alternative-gene naming data 1050, literature-recitation data 1055, and annotation data 1060. Many other examples are possible. Also, genomic data not currently available but that becomes available in the future may be accessed and locally maintained as described herein. Examples of remote databases 402 currently suitable for accessing in the manner described include GenBank, GenBank New, SwissProt, GenPept, DB EST, Unigene, PIR, Prosite, PFAM, Prodom, Blocks, PDB, PDBfinder, EC Enzyme, Kegg Pathway, Kegg Ligand, OMIM, OMIM Map, OMIM Allele, DB SNP, Gene Ontology, and PubMed. Hundreds of other databases currently exist that are suitable, and thus this list is merely illustrative.

[0082] Moreover, local genomic database 518 may also be supplemented with data obtained or deduced (by user-service manager 522) from other of the local databases serviced by database manager 512. In particular, although local products database 514 is shown for convenience of illustration as separate from database 518, it may be the same database. Alternatively, or all or part of the data in database 514 may be duplicated in, or accessible from, database 518.

[0083] More specific examples are now provided of how user service manager 522 may receive and respond to requests from user 101 for genomic information and for product information and/or ordering. These examples are described in relation to

Figures 7 through 12.

[0084] Figure 7 is a flow chart representing an illustrative method by which the illustrated embodiment of portal 400 may respond to a user's request for genomic or product information. In accordance with step 710 of this example, input manager 532 receives from client 410 over Internet 499 a request by user 101 for data. This request may, for instance, include an HTML, XML, or text document (e.g., tab delimited *.txt document) that includes user 101's selection of certain probe-set identifiers. As noted, the probe-set identifiers may be a number, name, accession number, symbol, graphical representation, or nucleotide or other sequence, as non-limiting examples. In some cases, user 101 may make this selection by employing one or more of analysis applications 199A to select probe-set identifiers (e.g., by drawing a loop around dots, as noted above) and then activating communication with portal 400 by any of a variety of known techniques such as right-clicking a mouse. The request may also, in accordance with any of a variety of known techniques, specify whether user 101 is interested in genomic and/or product data, as well as details regarding the type of data that is desired. For instance, user 101 may select categories of products, names of vendors or products, and so on from pull-down menus. Manager 532 provides user 101's request to user service manager 522, as described above.

[0085] In accordance with step 720, user-service manager 522 initiates an identification of user 101. Figure 8 is a block diagram showing the functional elements of manager 522 in greater detail, including account ID determiner 822 that, in this illustrative implementation, undertakes the task of identifying user 101. Determiner 822 may employ any of various known techniques to obtain this information, such as the use of cookies or the extraction from the user's request of an identification number entered by the user. Determiner 810, through database manager 512, may compare the user's identification with entries in user account database 519 to further identify user 101. In other implementations, the identity of user 101 need not be obtained, although statistics or information regarding user 101's request may be recorded, as noted above.

[0086] In accordance with step 725, user-service manager 522 formulates an appropriate query (using, for example, a version of the SQL language) for correlating probe-set

PCT/US2003/035559

identifiers with corresponding genes or EST's. Gene or EST determiner 820 is the functional element of manager 522 that illustratively executes this task. Determiner 820 forwards the query to database manager 512. If the probe-set identifiers provided by user 101 include sequence information, then the query may seek from database 511, and/or from SIF information in database 516, the identity of the one or more probe sets having a corresponding (e.g., similar in biological significance) sequence. If the included sequence information does not have a corresponding probe-set, such as a case in which user 101 has requested to have a probe-set produced, then determiner 820 may formulate an SQL statement, as in the above example, for the input of sequence information into the one or more appropriate databases. The sequence may be used as an identifier for an unknown, e.g., as yet not provided, probe-set. If the probe-set identifiers include one or more terms (e.g. referring to annotation information such as "tumor suppressor") then user service manager 522 identifies the genes, or EST's from database 518, where annotation information is stored with the corresponding genes or EST's. If the probe-set identifiers include names or numbers (e.g., accession numbers), then the query may seek the identity of the probe sets from database 516 that, as noted, includes data that associates names, numbers, and other probe-set identifiers with corresponding genes or EST's. User 101 may also have locally employed database application 230 to obtain this information, and included it in the information request in accordance with known techniques. In this case, step 725 need not be performed.

[0087]

As indicated in step 730, user-service manager 522 may then correlate the indicated genes and/or EST's with genomic information and/or product information. The performance of this task is undertaken by correlator 830 in the illustrated example. In one of many possible implementations, correlator 830 formulates a query via database manager 512 to database 513 in order to obtain links to appropriate information in local products database 514 and/or local genomic database 518. Figure 9 is a simplified graphical representation of database 513. Those of ordinary skill in the art will appreciate that this representation is provided for purposes of clarity of illustration, and that many other implementations are possible. In one aspect of an appropriate query to database 513, which is assumed for illustration to be a relational database, a gene or EST accession number 902 is associated with a link 904

to probe-set ID's 912. As indicated in Figure 9 by the association of both ID 902A and 902B to the same link 904N, multiple genes and/or EST's may be associated with the same probe-set ID. The information used to establish these associations is similar to that provided in database 516, as noted above, and the links may thus be predetermined or dynamically determined using database 516.

[0088] In other implementations, correlator 830 simply correlates one or more gene or EST identifiers, such as accession numbers, with products, such as biological products. These implementations are indicated in Figure 8 by the arrow directly from determiner 810 (which is optional) directly to correlator 830. The correlation may be accomplished according to any of a variety of conventional techniques, such as by providing a query to local products database 514, remote pages 404, and/or remote databases 402. These queries may be indexed or keyed by categories, types, names, or vendors of products, such as may be appropriate, for example, in examining look-up tables, relational databases, or other data structures. In addition, the query may, in accordance with techniques known to those of ordinary skill in the relevant art, search for products, product web pages, or other product data sources that are logically or syntactically associated with the gene or EST identifier(s). The results of the query may then be provided by output manager 534 to user 101, such as over Internet 499 to client 410.

[0089] A further implementation of correlator 830 is illustrated in figure 11, wherein cluster correlator 1100 receives from gene or EST determiner 820 a nucleotide sequence that may or may not correspond to a probe set. Cluster correlator then correlates the nucleotide sequence via database manager 512 with the corresponding protein sequence found in gene or EST to protein sequence data 1097, as is illustrated in figure 10. Alternatively, correlator 1100 may translate the nucleotide sequence into a protein sequence by methods known to those of ordinary skill in the art. Cluster correlator 1100 then sends the protein sequence to data storage and correlated data generators 1110, 1115, 1120, 1125, 1130, 1135, and 1140. The data storage and correlated data generators correspond to databases, now available or that may be developed in the future, that contain information regarding associated protein family, pathway, network, complex, and/or other protein annotation information. Such databases include but are not limited to, SCOP, Pfam, BLOCKS, EC, and GPCR, which

are known to those in the art as databases that contain annotation information. Such clusters of data may be stored in local genomic and/or product database 518 as illustrated in figure 10 as clustering data 1065, 1070, 1075, 1080, 1085, 1090, and 1095. The databases used in this example are for illustration only, and those of ordinary skill in the art know that many other examples are possible.

[0090] The data storage and correlated data generators use methods, known to those in the art as clustering methods, to determine sequence or structural similarity and alignments with similar protein sequences and/or structures. There are numerous types of clustering methods used for these purposes, for example what is commonly known as BLASTp represented in figures 10 and 11 as BLASTp clustering data 1085 and BLASTp data storage and correlated data generator 1130. Another example is commonly referred to as the Hidden Markov Model (referred to hereafter as HMM). HMM's are pattern matching algorithms that use a training set of data to "learn" the patterns contained in that training set of data. A preferred implementation is the so-called GRAPA set of HMM's that in the illustrated example are trained to be specific to families of proteins where each family has its own HMM trained to its characteristic pattern. A trained HMM can then analyze a sequence and return a score that corresponds to how well the sequence matches the pattern. In one illustrative implementation, a threshold value is assigned so that a score above the threshold is considered to be a member of the family and a score below is not. The data storage and correlated data generators of this implementation then generate what is commonly referred to as a pairwise alignment between the query sequence and the family consensus sequence, and correlate annotation data corresponding to the family.

[0091] Figure 12 is a representation of an illustrative example of a graphical user interface providing user 101 with information obtained by one clustering method. It will be appreciated by those of ordinary skill in the relevant art that numerous additional formats, both textual and graphical (e.g., connected node diagrams, dendograms, heat maps, and so on) may be used in other implementations. The clustering method used may be identified to user 101 by a graphical element such as illustrative element 1210. In this example, element 1210 includes the text "hmmpfam" that indicates that an HMM model representing an appropriate protein family

PCT/US2008/062559

characterized according to the Pfam database was employed. Graphical element 1240 is an identifier (e.g., "NP_003828") that identifies for user 101 a protein identified as the result of a query that user 101 posed and that resulted in the clustering results represented by data 1230. Further results of processing the user's query by employing the user-selected clustering method may also be displayed by graphical user interface 1200. For example, the user may be presented with one or more graphical elements such as returned family data 1220 and returned alignment data 1230. Family data 1220 displays the protein family that the query sequence belongs to as determined by the clustering method. A link to more detailed information is provided in this example. Alignment data 1230 displays the pairwise alignment between the query sequence and the family consensus sequence (although only a portion is shown for convenience in this illustration). Other graphical elements of GUI 1200 may include various other additional information that those of ordinary skill in the relevant art will appreciate to be useful in associating the query terms with biologically meaningful information and quantifying the degree of association. For example, the user may be presented with clustering scores derived by the clustering method, as illustratively indicated by expectation scores 1225. In another embodiment data processor 840 may return clustering results to user computer 100 and may also store them, or provide them for local processing, for use with probe-array analysis applications 199. As used herein, the term graphical user interface is intended to be broadly interpreted so as to include various ways of communicating information to, and obtaining information from, a user. For example, information may be sent to a user in an email as an alternative to, or in addition to, presenting the information on a computer screen employing graphical elements (such as shown illustratively in Figure 12). As is known by those of ordinary skill in the relevant art, the email may include graphics, or be designed to invoke graphics, similar to those that may be displayed in an interactive graphical user interface.

[0092]

One of many possible examples of the utility of these features includes a situation in which user 101 inputs a nucleotide sequence for which there is no corresponding probe set. The sequence is translated by correlator 830 into a protein sequence by known methods (alternatively, user 101 may have entered a protein sequence), and clustered using the HMM's for all, or any user-selected portion, of the available

databases. In the present example, it is assumed that a number of positive family identifications are made, and all related annotation data is presented to the user via a GUI as well as being stored on the user's LIMS system. After compiling and reviewing the annotation data, the user may choose to order a probe set that corresponds to the nucleotide sequence by including the new probe set in an order for a custom probe array.

[0093] In yet another example, a user may specify a sequence, which may for example be a putative gene, that does not correspond to any probe set. Correlator 830 correlates the user-specified sequence with one or more of the databases shown in Figure 10 as included in database 518, and identifies possibly related sequences (which may be related by family, functional, or other criteria other than, or in addition to, sequence). User-service manager 512 identifies the probe sets associated with the related sequences and/or the associated EST's, genes, and proteins. The identified probe sets, and optionally the array types in which they are represented, are provided to user 101 in an appropriate GUI and/or by other techniques such as email. Examples of probe-set annotations are provided in U.S. Provisional Patent Application, Serial No. 60/306,033, incorporated by reference above.

[0094] With respect to some specific implementations, one or more links 916 to related products and/or genomic data may be obtained by following the appropriate links to probe-set ID's 912. For example, link 904N may link to probe-set 912C, which is associated with links 916C to related product and/or genomic data. The information used to establish this association may be predetermined based on expert input and/or computer-implemented analysis (e.g., statistical and/or by an adaptive system such as a neural network) of the nature of inquiries by users. For example, it may be observed or anticipated (by humans or computers, as noted) that users conducting gene expression experiments resulting in the identification of certain genes may wish to use antibodies against the genes to conduct follow-on protein level experiments. The association between the genes and the appropriate antibodies may be stored in an appropriate database, such as database 516. Links 916C may thus include links to product or genomic data identifiers that identify links to data about the appropriate antibodies (for example, a link to product/genomic ID 922A), to catalogues of antibodies generally (e.g., ID 922B), or to a probe array specifically designed for

detecting alternatively spliced forms of the genes of interest (e.g., ID 922C). It is assumed for illustrative purposes that, in a particular aspect of this example, link 916C leads to ID 922C. Information about the availability of splice-variant probe arrays may be predetermined by the contents of links 926. For example, links 926D (associated with ID 922C, as shown) may be stored Internet and/or database-query URL's leading to vendor web pages, local products database 514, and/or local genomic database 518. Also, the content of links 926D may be dynamically determined by query of databases 514 or 518 or of remote data sources such as databases 402 or web pages 404. These and similar processes are represented by step 735 of Figure 7.

- [0095] As will now be appreciated by those of ordinary skill in the art, numerous variations and alternative implementations of this illustrative arrangement of database 513 are possible. For example, probe-set identification data may be linked to array identifiers (such as array ID 914), which may then be associated with links 916. As another of many possible examples, gene or EST accession numbers may be linked directly to product and/or genomic data ID 922 or, even more directly, to links 926. Implementations such as the illustrated one provide opportunities for making broad associations based on a more narrow inquiry by a user. For instance, a user may select only one probe-set identifier, but that identifier may be linked to multiple genes and/or EST's, which may be linked to multiple products or genomic data. In another example, link 926D may include a link to local genomic database 518. Based on the probe-set identifiers, gene or EST accession numbers, sequence information, or other data provided by or deduced from user 101's inquiry, database 518 may be searched for associated data in accordance with known query and/or search techniques.
- [0096] Returning now to Figure 7 and step 740 in particular, data returned in accordance with the query posed by correlator 830 is provided to either product data processor 842, genomic data processor 844, or both, as appropriate in view of the nature of the returned data. The functions of processors 842 and 844 are shown as separated for convenience of illustration, but it need not be so. Processors 842 and 844 apply any of a variety of known presentation or data transfer techniques to prepare graphical user interfaces, files for transfer, and other forms of data. This processed data is then provided to output manager 534 for transmission to client 410.

PCT/US2003/035559

[0097] In some implementations, user 101 may respond to the data thus transmitted by indicating a desire to purchase a product or receive further information. A request for further information may be processed in a manner similar to that described above with respect to Figure 7. If user 101 indicates a desire to purchase a product (see decision element 745), the indicated product may be prepared for shipment or otherwise processed, and the user's account may be adjusted, in accordance with known techniques for conducting e-commerce. As one of many alternative implementations, user-service manager 522 may notify the product vendor of user 101's order and the vendor may ship, or order the shipment of, the product. Manager 522 may then note, in one aspect of this implementation, that a fee should be collected from the vendor for the referral.

[0098]

In some implementations of portal 400, user 101 may provide to portal 400 (e.g., via client 410, Internet 499, and input manager 532) one or more gene or EST accession numbers or other gene or EST identifiers. Alternatively, or in addition, user 101 may provide to portal 400 one or more probe-set identifiers. User 101 may obtain the gene, EST, and/or probe-set identifier from a public source, from notations user 101 has taken as a result of experiments with a probe array or otherwise, from a list of genes or EST's having corresponding probes on a probe array, or from any other source or obtained in any other manner. Input manager 532 receives the one or more gene, EST, or probe-set identifiers and provides it or them to user-service manager 522, which formulates a query to database manager 512. In accordance with known query techniques and formats, the query seeks information from local products database 514 of product information related to the gene, EST, and/or probe-set identifiers. For this purpose, local products database 514 may be indexed, or otherwise searchable, for products based or keyed on any one or more of gene, EST, and/or probe-set identifiers. Some implementations may include, according to known techniques, similarity matching of a gene, EST, or probe-set identifier if, for example, all or part of a gene, EST, SIF (corresponding to the probe-set identifier) sequence is submitted. Also, a name-association function, in accordance with known techniques such as look-up tables, may be performed so that alternative names or forms of a gene, EST, or probe-set identifier may be found and used in the product data inquiry. In addition, in some implementations, manager 522 may initiate a remote data search

of remote databases 402 and/or remote vendor web pages 404, in accordance with known Internet search techniques, to obtain product information from remote sources. These searches may be based, for example, on product categories or vendors associated in local products database 514 with products, categories, or vendors associated with the gene, EST, or probe-set identifier provided by user 101. Manager 522 may provide product data corresponding to the gene, EST, and/or probe-set identifier, obtained from local products database 514 and/or remote pages or databases 404 or 402, and provide this product data to user 101 via output manager 534. For example, this product data may be included in web pages 524. In some of these implementations, portal 400 thus provides a system for providing product data, typically biological product data. The system includes input manager 532 that receives from user 101 one or more of a gene, EST, and/or probe-set identifier; user-service manager 522 that correlates the gene, EST, and/or probe-set identifier with one or more product data and that causes (e.g., via database manager 512) the product data to be obtained either locally from, e.g., database 514 or, in some implementations, remotely from, e.g., pages 404 or databases 402; and output manager 534 that provides the product data to user 101.

[0099] Similarly, a method is provided for providing biological product data, including the steps of: receiving from user 101 any one or more of a gene, EST, and/or probe-set identifier; correlating the gene, EST, and/or probe-set identifier with one or more product data; causing the product data to be obtained either locally from, e.g., database 514 and/or remotely from, e.g., pages 404 or databases 402; and providing the product data to user 101.

[0100] As indicated above, functional elements of portal 400 may be implemented in hardware, software, firmware, or any combination thereof. In the embodiment described above, it generally has been assumed for convenience that the functions of portal 400 are implemented in software. That is, the functional elements of the illustrated embodiment comprise sets of software instructions that cause the described functions to be performed. These software instructions may be programmed in any programming language, such as Java, Perl, C++, another high-level programming language, low-level languages, and any combination thereof. The functional elements of portal 400 may therefore be referred to as carrying out "a set

of genomic web portal instructions," and its functional elements may similarly be described as sets of genomic web portal instructions for execution by servers 510, 520, and 530.

- [0101] In some embodiments, a computer program product is described comprising a computer usable medium having control logic (computer software program, including program code) stored therein. The control logic, when executed by a processor, causes the processor to perform functions of portal 400 as described herein. In other embodiments, some such functions are implemented primarily in hardware using, for example, a hardware state machine. Implementation of the hardware state machine so as to perform the functions described herein will be apparent to those skilled in the relevant arts.
- [0102] Aspects of probe selection and design and other features applicable to implementations of the present invention are described in greater detail in the following patent applications, all of which are hereby incorporated by reference herein in their entireties for all purposes: U.S. Patent Applications Serial Nos. 10/028,884, titled "Method and Computer software Product for Genomic Alignment and Assessment of the Transcriptome," filed December 21, 2001; 10/027,682, titled Method and Computer Software Product for Defining Multiple Probe Selection Regions," filed December 21, 2001; 10/028,416, titled "Method and Computer Software Product for Predicting Polyadenylation Sites," filed December 21, 2001; and 10/006,174, titled "Methods and Computer for Designing Nucleic Acid Probe Arrays," filed December 4, 2001.
- [0103] Having described various embodiments and implementations, it should be apparent to those skilled in the relevant art that the foregoing is illustrative only and not limiting, having been presented by way of example only. Many other schemes for distributing functions among the various functional elements of the illustrated embodiment are possible. The functions of any element may be carried out in various ways in alternative embodiments. For example, some or all of the functions described as being carried out by determiner 820 could be carried out by correlator 830, or these functions could otherwise be distributed among other functional elements. Also, the functions of several elements may, in alternative embodiments, be carried out by

fewer, or a single, element. For example, the functions of determiner 820 and correlator 830 could be carried out by a single element in other implementations. Similarly, in some embodiments, any functional element may perform fewer, or different, operations than those described with respect to the illustrated embodiment. Also, functional elements shown as distinct for purposes of illustration may be incorporated within other functional elements in a particular implementation. For example, the division of functions between an application server and a network server of the genome portal is illustrative only. The functions performed by the two servers could be performed by a single server or other computing platform, distributed over more than two computer platforms, or otherwise distributed in accordance with various known computing techniques.

- [0104] Also, the sequencing of functions or portions of functions generally may be altered. Certain functional elements, files, data structures, and so on, may be described in the illustrated embodiments as located in system memory of a particular computer. In other embodiments, however, they may be located on, or distributed across, computer systems or other platforms that are co-located and/or remote from each other. For example, any one or more of data files or data structures described as co-located on and "local" to a server or other computer may be located in a computer system or systems remote from the server. In addition, it will be understood by those skilled in the relevant art that control and data flows between and among functional elements and various data structures may vary in many ways from the control and data flows described above or in documents incorporated by reference herein. More particularly, intermediary functional elements may direct control or data flows, and the functions of various elements may be combined, divided, or otherwise rearranged to allow parallel processing or for other reasons. Also, intermediate data structures or files may be used and various described data structures or files may be combined or otherwise arranged. Numerous other embodiments, and modifications thereof, are contemplated as falling within the scope of the present invention as defined by appended claims and equivalents thereto.

- [0105] What is claimed is: